BEHAVIOUR RECOGNITION AND EXPLANATION FOR VIDEO SURVEILLANCE

Neil Robertson[12], **Ian Reid**[2] **& Michael Brady**[2]

[1] QinetiQ, St Andrews Road, Malvern, UK
[2] University of Oxford, Dept. Engineering Science, Parks Road, Oxford, UK
{nmr,ian,jmb}@robots.ox.ac.uk

Keywords: Video surveillance, causal reasoning, Bayesian methods, action recognition.

Abstract

This paper is concerned with producing high-level reports and explanations of human activity in video from a single, static camera. The scenarios we focus on are urban surveillance and sports video where the imaged person is medium/low resolution. The final output is text descriptions which not only describe, in human-readable terms, *what* is happening but also *explain* the interactions which take place. The input to the reasoning process is the information obtained from lower-level algorithms which provide an abstraction from the image data to qualitative descriptions of human activity. Causal explanations of global scene activity, particularly where interesting events have occurred, is achieved using an extensible, rule-based method. The complete system represents a general technique for video understanding.

1 Introduction

A system which could automatically report on human activity in video would be extremely useful to surveillance officers who can be overwhelmed with increasingly large volumes of data. In both the civilian and military domains, the maintainence of situational awareness is critically important. This is due to the fact that, when an analyst focusses attention on a specific object of interest, potentially he/she is unaware of other interesting, suspicious or dangerous activity in the same scene. This problem is exacerbated when multiple screens must be monitored. Moreover, a system which could subsequently explain this activity would be a significant development in the technical area of video-based human behaviour understanding.

Computer Vision researchers, however, generally have focussed on developing lower-level techniques for analysing image sequences, such as feature-tracking, face/skin detection etc. Whereas the Artificial Intelligence community has contributed to the problem of expert knowledge representation and human-like reasoning processes. However, it has been recognised that there is a distinct lack of attempts to develop a system for visual scene understanding which combines the necessary aspects of both disciplines for intelligent visual surveillance. The work of this paper addresses this need without excluding the "man-in-the-loop". In fact, we utilise expert prior knowledge to ensure that the output descriptions on activity are accurate. To that end manual input is used to define the rules for the higher-level processes and to provide the training data labels. (In a simple urban surveillance scenario these qualitative descriptions might include, for example, *nearside-pavement*, *on the road*, *far-side pavement* for position, *left-to-right*, *away*, *towards* (etc.) for direction.) This work goes beyond simply reporting on individual activity, albeit at a human-readable level: in this paper we present a prototype system for *reasoning* about human activity in video. The system is split into two main parts: (i) low-level activity recognition for single agents in video which is described in section 3; (ii) higher-level reasoning about events using this information which is described in section 4. We demonstrate the efficacy of our system by presenting results from the urban surveillance domain (although the techniques are equally applicable to further applications e.g. sports footage as we show in other published work [10, 11]).

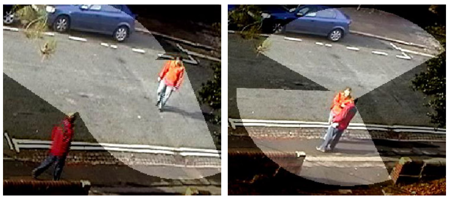


Figure 1: Gaze-direction is an important clue to intention.

The remainder of this paper is structured as follows. We begin with a review of the relevant prior art, then turn to a more detailed description of each of the stages of our algorithm. A technique to estimate where a person is looking is described in section 3.1. Single person spatio-temporal action recognition is described in section 3.2. Sequences of action comprise the overall behaviour of an individual, and we use HMMs to stochastically model these sequences in section 3.3. Together this qualitative information is defined as the information available to the "sensors" of a human agent (mainly, in this paper, a *pedestrian* agent). Rule-based reasoning, using expert knowledge about the domain, is introduced in section 4 to generate human-readable text explanations of the observed activity. The final result, in section 4, is therefore not only a high-level description of all scene activity but a causal explanation of interesting events. We conclude in section 5.

1.1 Related work

There has been much reported in the recent literature about methods for training recognition systems using large training data sets (e.g. [16]). Recently Zhong *et al.* [19] demonstrated detecting unusual activity by classifying motion and colour histograms into prototypes and using the distance from the clusters as a measure of novelty. Also Zelnik-Manor and Irani [18] used a distance metric to identify examples of actions in video. Boiman and Irani [2] address the problem of detecting "irregularities" in video, where "irregular" is defined solely by the context in which the video takes place. Xiang and Gong have addressed an important issue: how to effectively recognise action in a surveillance context when there is a sparsity of example data [17] and what rôle the high-level labelling of trajectories plays in this situation.

Making sense of a scene can be thought of as, "Assessing its potential for action, whether instigated by the agent or set in motion by forces already present in the world" [3]. In other words, a causal interpretation is most easily and most commonly judged by the motion effects that take place. Michotte, with Heider & Simmel demostrated that it is the kinematics of objects that produce the perception of causality, not appearance [13]. There is, nonetheless, a history in scene understanding research of analysing static scenes. In the work of Brand and Cooper [3]. One major shortfall in the reported work on reasoning, from an Artificial Intelligence perspective, is the lack of robust computer vision methods for obtaining low-level information about complex visual scenes and agents within them [9]. The work of Brand *et al.* relied on the extraction of very simple visual features from static images of blocks against a white background. Our work addresses this gap by applying established techniques to generate probabilistic estimates over qualitative descriptions of human activity in video [10, 11].

"Anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors" is an agent, according to Russell and Norvig [12]. An agent is, therefore, analogous to a software function. When human agents are combined, complex behaviour emerges which can model real-world behaviour as demonsrated by Andrade and Fisher for simulated crowd scenes [1]. There are many types of agent defined in the AI literature. The Belief-Desire-Intention agent is believed to model decision-making process humans use in every day life [7]. Related to agents, and of considerable relevance to the work of this paper, is the work of Dee and Hogg [5] in which a particular model of human behaviour is verified by comparing how "interesting" the model indicates the observed behaviour is to how worthy of further investigation a human believes the behaviour to be. Dee and Hogg's work focusses on inferring what an agent can sense through line-of-sight projection of rays and the subsequent use of a predefined model of goal-directed behaviour to predict how the agent is expected to behave. Not all of the information required for reasoning is automatically extracted from the images (which is an area we explicitly address in this work).

On rule-based reasoning, Siler notes that rules have, "... shown the greatest flexibility and similarity to human thought processes ..." [15]. These rules can quickly be identified and written down by an expert. A significant positive aspect of rule-based reasoning is that it is easy to update the system's knowledge by adding new rules without changing the reasoning engine [9]. It is also easy to transfer between applications by specifying a new set of rules.

2 Contributions

In relation to the prior work in this area, the contributions of this paper are:

- Our method requires much less training data than the statistical learning techniques found in the literature (hours vs. days),
- We explicitly use the prior knowledge of the expert analyst which is not only technically advantageous (i.e. provides more accurate results), but is strongly aligned with the needs of surveillance professionals,
- Our system achieves reasoning about causal relations between human agents direct from an input video using complex visual features, which has not been demonstrated until now.

3 Low-level activity estimation direct from video

As we have stated, the overall goal is that we may be able to automatically reason about human activity in video. The information we require to achieve this goal becomes apparent when we consider what a human might need to know to *reason* about human activity e.g. What are the agents doing? What can the agents sense?

In contrast to the previous work in this area, the low-level vision techniques we have developed and proven in earlier published work answer these questions automatically in a fully probabilistic (Bayesian) fashion [10, 11]. Probability distributions of the following information is extracted direct from the video: Gaze-direction, Spatio-temporal action and Behaviour (which is modelled as a sequence of spatio-temporal actions).

This is done using our existing lower-level vision algorithms. This probabilistic data is then used as input to a deterministic rule-based reasoning engine. These activity descriptions

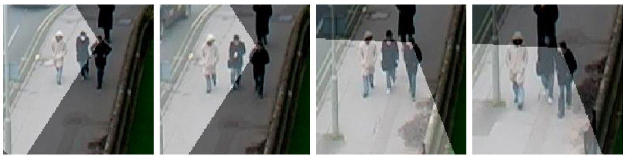


Figure 2: In surveillance scenarios, groups of people together can be identified using gaze-direction.

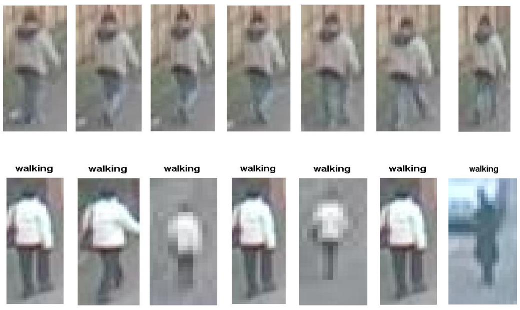


Figure 3: Action recognition is achieved using optic-flow based descriptors.

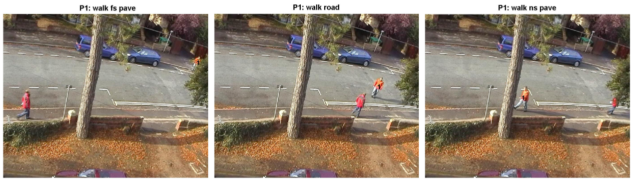
taken on their own comprise a *report* of the video. We then use this report to *explain* observed interactions using a rule-based reasoning approach. In this section we describe the probabilistic activity estmation: gaze-direction estimation (section 3.1); spatio-temporal action recognition (section 3.2); and finally behaviour recognition (section 3.3).

3.1 Gaze direction estimation

The first lower-level component of our system estimates where a person is looking in images where the head is typically in the range 20 to 40 pixels high [11]. We use a feature vector based on skin detection to estimate the orientation of the head, which is discretised into 8 different orientations, relative to the camera. A fast sampling method returns a distribution over previously-seen head-poses. The overall body pose relative to the camera frame is approximated using the velocity of the body, obtained via automatically-initiated colour-based tracking in the image sequence [4]. By combining direction and head-pose information gaze is determined more robustly than using each feature alone. We show examples of this process applied to surveillance footage in figures 1 and 2.

3.2 Action and Behaviour recognition

In addition to gaze-direction we also require to extract basic information such as position, velocity and activity-type e.g. walking vs. running vs. standing etc. To that end we employ a technique for sampling from hand-labelled exemplar databases [14]. This sampling method returns a probability distribution over a set of training examples, where the qualitative labels of place, direction and action-type have been identified by an expert user. This method holds three significant advantages: (i) high-level descriptions can be incorporated by a qualified expert; (ii) by sampling non-parametrically from the data, far less training data is required than is the case for standard, statistic-based learning techniques such as HMMs; (iii) probabilistic distributions prevent us committing to one interpretation of activity too early.



Frame	Activity	Likelihood
1 - 70	Walking on far-side pavement	0.86
71 - 225	Walking on road	0.94
226 - 450	Walking on near-side pavement	0.94

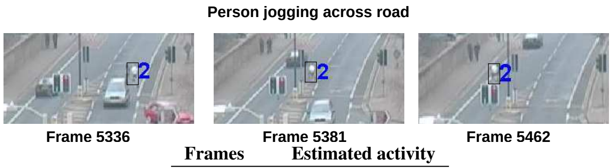
Figure 4: An accurate commentary is obtained for this urban street scene where the person moving in from the top-right of the images is under observation.

Position and velocity exemplars are derived directly from the centroid of the object as estimated using a colour-based tracker [4]. Action-type is encoded using a descriptor based on optic-flow, which is an extension to the descriptor of Efros *et al.* [6]. An example of matching actions using this method is shown in figure 3.

The position, velocity and action-type databases are maintained independently. This enables more efficient use of each feature, reducing the volume of training data required. Bayesian fusion allows us to compute probability distributions over *spatio-temporal* actions (such as "walking on near-side pavement") from the independent distributions over the feature databases. By taking the ML estimate from this distribution over all possible spatio-temporal actions at each time step, a commentary on activity is generated. An example of this for surveillance video is shown in figure 4. The priors on spatio-temporal actions can be derived directly from the training datasets, on the basis of frequency of occurrence, or can be easily hand-tuned. In the commentary example of figure 5, the priors are critical to the choice of the correct spatio-temporal action. Running is not represented as often in the example database. Therefore if the priors for each simple-action are computed on the basis of frequency then the ML spatio-temporal action for this sequence is *road, walking*. If however, the priors are uniform the ML result is as shown. Note that in either case the correct activity is still represented in the distribution over spatio-temporal actions.

3.3 Behaviour as a sequence of spatio-temporal actions

Having generated probability distributions over actions, we subsequently use Hidden Markov Models to encode known



Frames	Estimated activity
5330-5470	road, running

Figure 5: A second example from a more challenging surveillance scene.

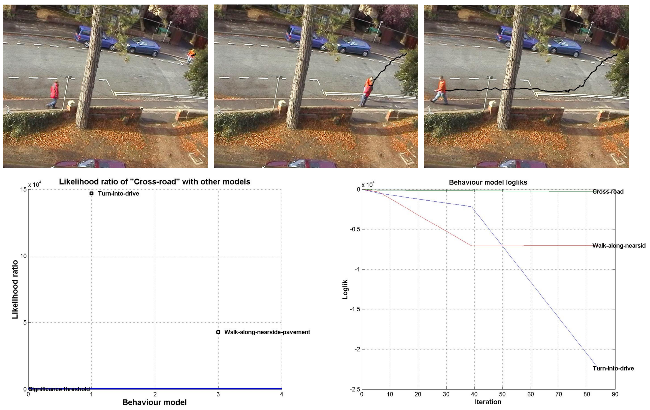


Figure 6: For the sequence in the top row we compute (*left*) the likelihood ratio of the most likely model with the other behaviour models in the bank of models. The likelihood of the behaviour model HMMs over the entire sequence is shown (*right*).

rules about behaviour. The ML spatio-temporal action is an abstraction from the images themselves to a description of activity in the scene in general. Taken on its own, it provides a report, or commentary, on activity. It is, therefore, not dependent on one particular camera viewpoint. This enables us to derive the action sequence from an automatic parse of behaviour. The hidden state of the HMM corresponds to a distribution over spatio-temporal actions. For the scene in figure 6 we encoded 3 such HMM behaviour models very efficiently ("crossing road", "walking along pavement" and "turning into drive") by defining the transition and initial-state probabilities for each model. On-line estimation of which model best explains the observed ML action-sequence (not the image data) enables us to estimate higher-level behaviour. The Likelihood Ratio is used for model-selection. Note that, even if the global behaviour is not recognised a sensible description of activity can still be achieved from the action-recognition stage of our system. Also, since these HMM behaviour models are general to the scene, they can discriminate between the same type of behaviour performed in different ways without the need for separate models (as a learning technique trained directly from the image data would require). An example of this feature in operation is shown in figure 7.

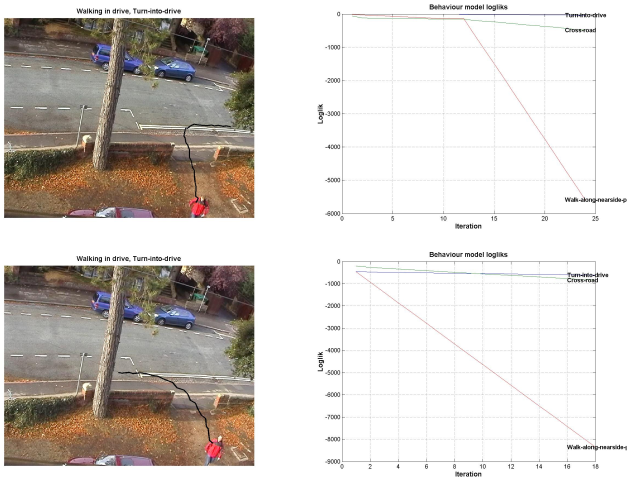


Figure 7: A single HMM associated with the *turning-into-drive* behaviour is used to classify the same behaviour but performed in different ways.

4 High-level causal reasoning about interesting activity

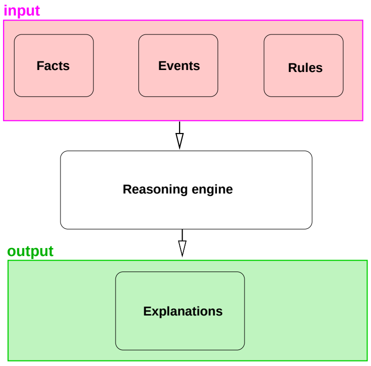


Figure 8: This diagram outlines the reasoning process we use for explaining activity in video. "Facts" are derived directly from video, "events" and "rules" are hand-coded for a particular scenario.

Having automatically extracted human-readable descriptions of action, behaviour and gaze-direction for pedestrian agents in video, we are now in a position efficiently to encode a reasoning process to explain "interesting" activity. The overall process is based on predefined rules and is shown in figure 8. A set of "facts", derived from the application of the low-level activity recognition algorithms (described in section 3) to the input video stream, is maintained. This comprises all that is known about the agent's activity. For a particular scene, certain "trigger events" which require explanation are predefined manually, as are known rules about normal human behaviour for the scene. These can be encoded at a high-level only because we automatically derive qualitative human-readable descriptions of activity. The same reasoning engine which is used across video sequences from the same and very different application domains. (Although the trigger events and the rules require updating for different scenarios.) As an illustrative example, in an urban surveillance scenario the event “move-to-road” is generated by a transition between the actions *walking-on-farside-pavement* and *walking-on-road*. Intermediate events such as “meeting” or “ignoring” are inferred using rules which utilise all available information (including gaze direction). The hypothetical explanations for the activity are defined as follows:

1. IF the event “move-to-road” is followed by event “move-to-pavement” AND the current location is not the same as the location triggering the first event (i.e. the road is crossed) AND, subsequently, a meeting takes place THEN the explanation is that, “the agent crossed the road to meet the other agent”,

2. IF a crossing of the road is observed NOT followed by an interaction THEN the explanation is that the agent crossed the road,

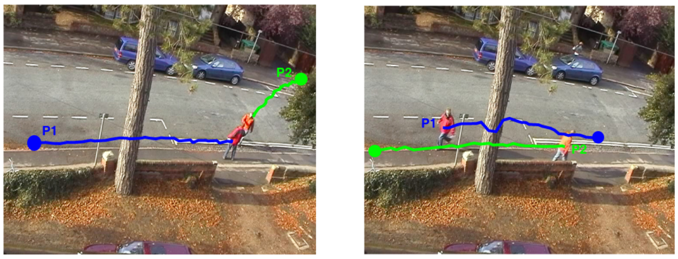
3. IF a “move-to-road” event is triggered AND subsequently a “move-to-pavement” event but back to the same pavement THEN no explanation is provided UNLESS another agent was in the near vicinity THEN the explanation is that it was necessary to avoid collision.

To demonstrate how extracting qualitative, intermediate descriptions of activity aids the encoding of rules, the pseudo-code for the reasoning process initiated by the scenario “move-to-road” is shown in algorithm 1. Similarly, we can generate hypotheses for explaining events such as “stopping”, “move-to-pavement” and “move-to-driveway”. It can be seen that the rule-set is (a) general to all such urban scenes, (b) easily augmented (i.e. by adding more rules). In figure 9, the output for two different situations, automatically generated by our system, is shown. Exactly the same engine and events-set is applied to the urban scene shown in figure 10. The rules are augmented with knowledge that the road may legitimately be crossed at the pedestrian crossing i.e. despite there being no evidence for a meeting, crossing at the lights is a plausible reason for the observed behaviour.

Finally, for interest and to demonstrate the generality of our reasoning process, in figure 11 we show an automatic explanation of traffic activity. In this case, the input activity description is limited (by comparison to the main results of this paper) but nonetheless demonstrates the utility of a rule-based reasoning system when an intermediate, qualitative estimation of low-level motion has been achieved.

Algorithm 1 move-to-road rule

```
load facts
if event=“meeting” then
  for $j = 1$ to lastFrame do
    if scenario = “meeting” then
      currentAction = facts.positionLabel(j)
      explanation = “Person” event “to meet on” currentAction
    end if
  end for
  for $j = 1$ to lastFrame do
    if scenario = “ignore” then
      currentAction = facts.positionLabel(j)
      explanation = “Person” event “to avoid other Person on” currentAction
    end if
  end for
end if
```



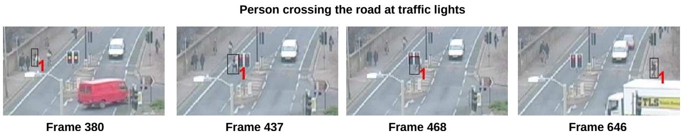
Explanation	Explanation
P2 move-to-road to Meet on ns-pavement	P1 move-to-road to Avoid P2 on ns-pavement
P2 move-to-pavement to Get-off-road	P1 move-to-pavement to Get-off-road

Figure 9: Causal explanations of interactions in an urban scene are automatically generated.

5 Conclusion

In this paper, we have presented a complete system for generating high-level commentary on human activity in video and for reasoning, causally, about interesting events. We began by posing the question *What does an agent require to know in order to reason about a scene?* To answer this question we have exploited recent developments in Computer Vision with regard to action and behaviour recognition. The information we extracted was sufficient to enable not only the generation of accurate, human-readable commentary on surveillance video, but also (and most significantly) causal explanations of interesting activity. This is the first demonstration of such a system which is (a) general for video sequences where the imaged person is low/medium resolution, and (b) complete, operating directly from the video stream to generate explanations of events, while utilising the “man-in-the-loop” and using complex visual features.

The most pressing area for further development is to demonstrate fully probabilistic reasoning. For reasons of expediency, we have used the ML result from the vision components of our system, but we suggest that a Bayesian Network might equally well allow causal relationships to be inferred while retaining the benefits of probabilistic models



Commentary of activity	Explanation
NE-pavement, walking	Person move-to-road to Cross-road at N-ped-crossing
N-ped-crossing, walking	Person move-to-pavement to Get-off-road
N-ped-crossing, stopped	
N-ped-crossing, walking	
NW-pavement, walking	

Figure 10: The same rules and events set as used to generate the results in figure 9 is successfully used here in a different scene.

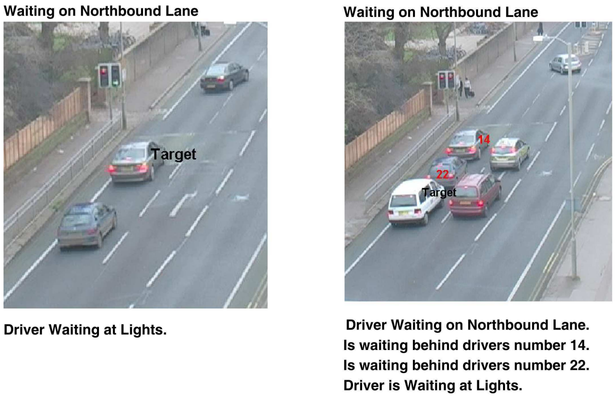

Waiting on Northbound Lane
Target
Driver Waiting at Lights.
Waiting on Northbound Lane
14
22
Target
Driver Waiting on Northbound Lane.
Is waiting behind drivers number 14.
Is waiting behind drivers number 22.
Driver is Waiting at Lights.

Figure 11: Resolution of potential anomalies i.e. why did the car stop? (*left*) and understanding of queues (*right*) is achieved using our method.

(such as preventing committing to one decision too early in the reasoning chain). Pearl's work on Causality is likely to be relevant to this problem [8].

References

[1] E.L. Andrade, R.B. Fisher *Simulation of Crowd Problems for Computer Vision* First International Workshop on Crowd Simulation (V-CROWDS '05), Lausanne, Nov 2005

[2] O. Boiman and M. Irani *Detecting Irregularities in Images and in Video* IEEE International Conference on Computer Vision (ICCV), Beijing, October 2005.

[3] M.Brand, L.Birnbaum, P.Cooper *Sensible scenes: visual understanding of complex structures through causal analysis*, Proceedings, National Conference on Artificial Intelligence, Washington D.C., 1993.

[4] D. Comaniciu and P. Meer *Mean Shift Analysis and Applications* Proceedings of the International Conference on Computer Vision-Volume 2, p.1197, September 20-25, 1999

[5] H. Dee and D. Hogg *Detecting Inexplicable Behaviour* Proceedings of the British Machine Vision Conference, 2004

[6] A.A. Efros, A. Berg, G. Mori and J. Malik *Recognising Action at a Distance* Proceedings of the International Conference on Computer Vision, Nice, France, July 2003

[7] M. Georgeff, B. Pell, M. Pollack, M. Tambe, M. Wooldridge *The Belief-Desire-Intention Model of Agency* Proceedings of the 5th International Workshop on Intelligent Agents V : Agent Theories, Architectures, and Languages (ATAL-98)

[8] J. Pearl *Causality. Models, Reasoning and Inference* Cambridge University Press, 2000, ISBN 0 521 77362 8

[9] M. Rigolli, *D.Phil. Thesis*, Department of Engineering Science, University of Oxford, 2006.

[10] N.M. Robertson and I.D. Reid *Behaviour understanding in video: a combined method* Proceedings of the International Conference on Computer Vision (ICCV), October 2005, Beijing, China

[11] N.M. Robertson and I.D. Reid *Estimating Gaze Direction from Low-Resolution Faces in Video* Proceedings of the 9th European Conference on Computer Vision (ECCV), Graz, Austria, May 2006

[12] S. Russel and P. Norvig *Artificial intelligence, a modern approach* Prentice-Hall, 1995

[13] B.J. Scholl *Innateness and (Bayesian) visual perception* In P. Carruthers, S. Laurence and S. Stich (Eds.), The innate mind: Structure and contents (pp. 34 - 52). Oxford University Press, 2005

[14] H. Sidenbladh M. Black, L. Sigal. *Implicit Probabilistic Models of Human Motion for Synthesis and Tracking* European Conference on Computer Vision, Copenhagen, Denmark, June 2002.

[15] *Fuzzy Expert Systems and Fuzzy Reasoning William Siler* James J. Buckley ISBN: 0-471-38859-9, January 2005

[16] P. Viola, M. Jones, D. Snow *Detecting Pedestrians using Patterns of Motion and Appearance* Proceedings of the International Conference on Computer Vision, Nice, France, July 2003 Pub. Morgan Kaufmann, Palo Alto, CA, USA, 1990.

[17] T. Xiang and S. Gong *Video behaviour profiling and abnormality detection without manual labelling* In Proc. International Conference on Computer Vision (ICCV), Beijing, China, October 2005.

[18] L. Zelnik-Manor and M. Irani *Event-Based Video Analysis* IEEE Conference on Computer Vision and Pattern Recognition (CVPR), December 2001

[19] H. Zhong, J. Shi and M. Visontai *Detecting Unusual Activity in Video* Computer Vision and Pattern Recognition, Washington D.C., USA, June 2004